Notice of Allowability	Application No.	Applicant(s)	
	10/647,203	FRANZ ET AL.	
	Examiner	Art Unit	
	PARAS SHAH	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *01/14/2009*.

2. ☒ The allowed claim(s) is/are *55-69*.

3. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

4. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☒ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☐ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application
6. ☒ Interview Summary (PTO-413), Paper No./Mail Date _____ .
7. ☒ Examiner's Amendment/Comment
8. ☒ Examiner's Statement of Reasons for Allowance
9. ☐ Other _____.

DETAILED ACTION

1. This communication is in response to the Arguments and Amendments filed on 01/14/2009. Claims 55-69 remain pending and have been examined. The Applicants' amendment and remarks have been carefully considered and they have been found to be persuasive. Accordingly, this application is in condition for Allowance.

2. All previous objections and rejections directed to the Applicant's disclosure and claims not discussed in this Office Action have been withdrawn by the Examiner.

EXAMINER'S AMENDMENT

3. An examiner's amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 CFR 1.312. To ensure consideration of such an amendment, it MUST be submitted no later than the payment of the issue fee.

Authorization for this examiner's amendment was given in a telephone interview with Dan Burns and Xin Ma on 02/02/2009.

The application has been amended as follows:

Claims: **Replace Claim 55** from " A computer-implemented method for identifying compounds in text, comprising: extracting a vocabulary of tokens from text; iterating from n > 2 down to n = 2 where n decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each n-gram into n-1 pairs of two adjacent segments,

where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each pair of segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation; identifying a set of n-grams having scores above a threshold; and adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary." **To** -- A computer-implemented method for identifying compounds in text, comprising: extracting a vocabulary of tokens from text; iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each n-gram into n-1 pairs of two adjacent segments, where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each pair of the n-1 pairs of two adjacent segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation for the each of the n-1 pairs; identifying a set of n-grams having scores above a threshold; and adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary, wherein the iterating is performed by one or more processors.—

Replace Claim 60 from "A storage device storing program code, which, when executed by a processor, causes the processor to perform operations comprising: extracting a vocabulary of tokens from text; iterating from n > 2 down to n = 2 where n

decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each n-gram into n-1 pairs of two adjacent segments, where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each pair of segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation; identifying a set of n-grams having scores above a threshold; and adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary." **To** -- A computer readable storage medium on which program code is stored, which program code, when executed by a processor, causes the processor to perform operations comprising: extracting a vocabulary of tokens from text; iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each *n*-gram into *n*-1 pairs of two adjacent segments, where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each of the *n*-1 pairs of two adjacent segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation for the each of the *n*-1 pairs; identifying a set of n-grams having scores above a threshold; and adding the identified set of n-grams as compound tokens to the

vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary.—

Replace Claim 61 from "The storage device of claim 60...." **to** --The computer-readable storage medium of claim 60....—

Replace Claim 62 from "The storage device of claim 61...." **to** --The computer-readable storage medium of claim 61....—

Replace Claim 63 from "The storage device of claim 61...." **to** --The computer-readable storage medium of claim 61....—

Replace Claim 64 from "The storage device of claim 60...." **to** --The computer-readable storage medium of claim 60....—

Replace Claim 65 from "A system comprising: a computer readable medium including a program product; and one or more processors configured to execute the program product and perform operations comprising: extracting a vocabulary of tokens from text; iterating from n > 2 down to n = 2 where n decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each n-gram into n-1 pairs of two adjacent segments, where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each pair of segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation; identifying a set of n-grams having scores above a threshold; and

adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary." **To** -- A system comprising: a computer readable storage medium on which a program product is stored; and one or more processors configured to execute the program product and perform operations comprising: extracting a vocabulary of tokens from text; iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of: identifying a plurality of unique n-grams in the text, each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary; dividing each n-gram into n-1 pairs of two adjacent segments, where each segment consists of at least one token; for each n-gram, calculating a likelihood of collocation for each of the n-1 pairs of two adjacent segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation for the each of the n-1 pairs; identifying a set of n-grams having scores above a threshold; and adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary.—

Reasons for Allowance

4. Claims 55-69 are allowed

5. The following is an examiner's statement of reasons for allowance:

The closest prior arts of record with respect to independent claims 55, 60. and 65, Su *et al.* (In *Proceedings of the 32nd Annual Meeting on Association For*

Computational Linguistics 1994) is cited to disclose extracting a vocabulary (see page 244, 2nd full paragraph, sect. Simulation, (1st paragraph), line 5-8, compound list) of tokens(see page 244, Table 1) from text (see page 243, left column, 2nd paragraph, line 6);identifying a plurality of unique n-grams in the text (see page 245, right column, "Simulation," 1st paragraph, compound list is modified or rebuild after a new compound word is detected.), each n-gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary (see page 244, left column, lines 4-18, relative frequency of the n-gram is computed.); identifying a set of n-grams having scores above a threshold (see page 243, right column, line 23); and adding the identified set of n-grams as compound tokens to the vocabulary (see page 245, right column, 2nd paragraph, line 7, compound list) and removing constituent tokens that occur in the added compound tokens from the vocabulary (see page 244, left column, Relative Frequency Count paragraph). However, Su *et al.* does not specifically disclose dividing each n-gram into *n*-1 pairs of two adjacent segments, calculating a likelihood of collocation for each of the *n*-1 pairs of two adjacent segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation for the each of the *n*-1 pairs. The combination of the limitations stated above are not taught or suggested by Su.

Frantzi *et al.* ("Extracting Nested Collocations") is cited to disclose the use iterating from n > 2 down to n = 2 where n decreases by one each iteration and in each iteration performing the actions (page 43, right column, "The algorithm …", 2nd full paragraph, code underneath and page 44, entire left column-right column, numbered

item 5) (e.g. From the cited reference it is seen that the n-gram starts from some maximum limit and then proceeds to a lower order n-gram. The n-gram is decremented and takes into account the frequency of occurrence in order to determine a candidate collocation by the determination of a C value.) However, Frantzi et al. does not specifically disclose division of each n-gram into n-1 pairs of two adjacent segments and the likelihood of collocation determined for each of the n-1 pairs of two adjacent segments based on a lowest calculated likelihood of collocation for each of the n-1 pairs.

Thus, independent claims 55, 60, and 65 are allowable over the prior art of record because the cited prior art alone or in combination, does not fairly suggest or disclose the claimed features, in combination, which have been mentioned above in the prior arts of record.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably accompany the issue fee. Such submissions should be clearly labeled "Comments on Statement of Reasons for Allowance."

Conclusion

6. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

Light (US 5,8421,217) is cited to disclose recognition of compound words in documents. Sassano (US 5,867,812) is cited to disclose a compound-word dictionary fir

determining and adding compounds. Smadja (US 6,173,298) is cited to disclose a dynamic collocation dictionary based on bigrams. Ejerhed (US 6,754,617) is cited to disclose determination of solid compound words. Kaku et al. (US 2007/0067157) is cited to disclose extraction of interesting phrases and adding to a dictionary.

7. Any inquiry concerning this communication or earlier communications from the examiner should be directed to PARAS SHAH whose telephone number is (571)270-1650. The examiner can normally be reached on MON.-THURS. 7:00a.m.-4:00p.m. EST.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Patrick Edouard can be reached on (571)272-7603. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/P. S./
Examiner, Art Unit 2626

02/03/2009
/Patrick N. Edouard/
Supervisory Patent Examiner, Art Unit 2626